

# A comparison of some criteria for states selection in the latent Markov model for longitudinal data

Silvia Bacci\*, Silvia Pandolfi\*<sup>†</sup>, Fulvia Pennoni<sup>‡</sup>

December 4, 2012

## Abstract

We compare different selection criteria to choose the number of latent states of a multivariate latent Markov model for longitudinal data. This model is based on an underlying Markov chain to represent the evolution of a latent characteristic of a group of individuals over time. Then, the response variables observed at the different occasions are assumed to be conditionally independent given this chain. Maximum likelihood of the model is carried out through an Expectation-Maximization algorithm based on forward-backward recursions which are well known in the hidden Markov literature for time series. The selection criteria we consider in our comparison are based on penalized versions of the maximum log-likelihood or on the posterior probabilities of belonging to each latent state, that is the conditional

---

\*Department of Economics, Finance and Statistics, University of Perugia, Italy; *email*:  
silvia.bacci@stat.unipg.it.

<sup>†</sup>*email*: pandolfi@stat.unipg.it

<sup>‡</sup>Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milan, Italy; *email*:  
fulvia.pennoni@unimib.it

probability of the latent state given the observed data. A Monte Carlo simulation study shows that the indices referred to the log-likelihood based information criteria perform in general better with respect to those referred to the classification based criteria. This is due to the fact that the latter tend to underestimate the true number of latent states, especially in the univariate case.

**Keywords:** Akaike Information Criterion Bayesian Information Criterion entropy mixture model multivariate latent Markov model Normalized Entropy Criterion.

## 1 Introduction

A crucial element in the literature about the wide class of mixture models (McLachlan and Peel, 2000) is represented by the choice of the number of mixture components, which represents a specific aspect related with the model selection process. For instance, this issue arises in the context of latent class (LC) models about the choice of latent classes and in the contexts of hidden Markov (HM) models for time-series and stochastic processes (Zucchini and MacDonald, 2009) and of latent Markov (LM) models (Wiggins, 1973) for longitudinal data. This last class of models is typically used when the interest is in describing the evolution of a latent characteristic of a group of individuals over time. They assume that one or more occasion-specific response variables depend only on a discrete latent variable, characterized by a given number of latent states, which follows a first-order Markov process (Bartolucci et al, 2013). The basic idea behind this assumption is that the latent process fully explains the observable behavior of a subject. Furthermore, the latent state to which a subject belongs to at a certain occasion only depends on the

latent state at the previous occasion. An LM model may also be seen as an extension of the LC model, in which the assumption that each subject belongs to the same latent class throughout the period of observation is suitable relaxed.

In such a context the number of latent states is usually selected on the basis of the observed data, both in the case of the basic LM model or in the advanced versions that, for example, allow for the inclusion of observable individual covariates. Only in certain applications the number of latent states is a priori defined by the nature of the problem or by the interest of the research. However, states selection on the basis of the observed data implies that increasing the number of states often improves the fit of the model, as judged by the likelihood, but also the number of parameters. The same problem arises when selecting the number of components in a finite mixture model.

The more common approaches which have been adopted to balance model fit and parsimony are based on information criteria constructed according to indices that are penalized versions of the maximum log-likelihood. Among these criteria, the most common are the Akaike Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978). The first one is known as an estimator of the Kullback-Leibler discrepancy between the model generating the data and the fitted model. BIC may be instead seen as an asymptotic approximation of the integrated likelihood, which provides an estimator of a transformation of the Bayesian posterior probability of a candidate model. Several alternative to the AIC criterion have been proposed in literature such as  $AIC_3$  of Bozdogan (1993) and the Consistent AIC (CAIC) criterion proposed by Bozdogan (1987) which are based on different penalization terms. It is important to mention that the information criteria are preferred to methods based on the likelihood ratio

test between nested models because the latter require bootstrap resampling procedure.

In addition to the above log-likelihood based information criteria, classification based criteria have been proposed in literature, which allow us to measure the quality of the classification provided by a model. The Normalized Entropy Criterion (NEC) is an approach first developed by Celeux and Soromenho (1996) to select the number of components in the context of mixture models. It is based on an entropy term computed on the basis of the posterior probabilities for every sample unit and mixture component. This criterion takes into account the quality of the classification, and then how well the clusters are separated, further to the goodness-of-fit of the model, which is measured in terms of log-likelihood. An entropy index has been recently proposed in HM literature to measure uncertainty involved in connection with finding the most likely sequence of the latent states; see Hernando et al (2005) and Durand and Guédon (2012). The entropy measure, however, has not been investigated as a tool for states selection in such a context. Among other classification based criteria, it is worth mentioning also the Classification Likelihood information Criterion (CLC), adopted by Biernacki and Govaert (1997) in the mixture context, and an approximation of the Integrated Classification Likelihood criterion (ICL; Biernacki et al, 1998) using BIC denoted as (ICL-BIC) firstly adopted by Biernacki et al (2000) and McLachlan and Peel (2000).

In the context of finite mixture models and, in particular, of LC models, several studies exist aimed at comparing the performance of the above mentioned criteria. Among others, Fraley and Raftery (2002) used BIC for clustering in mixture models, showing its satisfactory behavior (see also McLachlan and Peel, 2000, Ch. 8). Simulation studies have also been performed by Nylund et al (2006) for growth mixture and LC models,

and by Biernacki and Govaert (1999) for Gaussian mixtures. In both situations it was found that BIC outperforms the other information criteria. We also refer to Dias (2006) for a study refer to the LC model with binary response variables in which emerges that  $AIC_3$  is the best criterion for selecting the number of latent classes. Moreover, CAIC has been proved to have a similar performance with respect to BIC (Lin and Dayton, 1997). About the behavior of the classification based criteria, we refer to Biernacki and Govaert (1999), which found that NEC gives poor results in selecting the model under comparison, although it exhibits good behavior in detecting the number of clusters. Moreover, Biernacki et al (2000) showed that ICL appears to be more robust than BIC to violation of some of the mixture model assumptions.

Even if these criteria are widely used in literature, their performance have not been studied enough in detail in connection with LM models. A comparison of AIC and BIC performance in connection with states selection of a univariate LM model may be found in Bartolucci et al (2013)[Ch. 7]. However, to our knowledge, there are no studies aimed at comparing the behavior of the different information criteria mentioned above. On the other hand their properties have been studied in the context of HM models; see, among others, Celeux and Durand (2008), Costa and De Angelis (2010) and the references therein. However, the context is quite different, since HM models are used for time series, whereas LM models are applied to longitudinal data. The main purpose of this paper is to compare the performance of all the illustrated information criteria when applied to select the number of latent states in a multivariate LM model.

We show a Monte Carlo simulation study on the basis of different model specifications, with respect to the number of response variables, and to different conditional response

probabilities and transition probabilities. The aim is to analyze the effect of these factors on selecting the number of states and to set up a comparison between log-likelihood based and classification based criteria. In particular, in applying the NEC criterion, we consider an entropy measure based on the posterior probabilities of all the possible configurations of latent states, given the observed data, for every sample unit. We also consider two approximations of NEC which are based on a modified version of the entropy computed on the basis of the posterior probability of every single latent state at every time occasion.

The article is organized as follows. In the following section we illustrate the multivariate LM model and we deal with maximum likelihood estimation of the model parameters. In Section 3 we illustrate the latent states selection criteria under comparison. In Section 4 we show the results of a series of simulations made in order to assess the quality of the analyzed criteria. In Section 5 we provide main conclusions.

## 2 The multivariate LM model

In the multivariate formulation of the LM model (that includes the univariate one as a special case) we observe a vector of categorical response variables  $\mathbf{Y}^{(t)} = (Y_1^{(t)}, \dots, Y_r^{(t)})$ , for  $t = 1, \dots, T$ . Each variable  $Y_j^{(t)}$ ,  $j = 1, \dots, r$ , has  $c_j$  categories, labeled from 0 to  $c_j - 1$ . We denote by  $\mathbf{Y} = (\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(T)})$  the vector of observed responses made of the union of vectors  $\mathbf{Y}^{(t)}$ , which usually, is referred to repeated measurements of the same variables  $Y_j$  ( $j = 1, \dots, r$ ) on the same individuals at different time points.

The model is based on two main assumptions. Firstly, the vectors  $\mathbf{Y}^{(t)}$  are conditionally independent given a latent process  $\mathbf{U} = (U^{(1)}, \dots, U^{(T)})$ , and the response variables

in each of these vectors are conditionally independent given  $U^{(t)}$  at time  $t$  with state space  $\{1, \dots, k\}$ . In other words, each occasion-specific observed variable  $Y_j^{(t)}$  is independent of  $Y_j^{(t-1)}, \dots, Y_j^{(1)}$  and of each  $Y_h^{(t)}$ , for all  $h \neq j = 1, \dots, r$ , given  $U^{(t)}$ . This is the so called *local independence* assumption. Secondly, the latent process  $\mathbf{U}$  is assumed to follow a first-order Markov chain with  $k$  latent states, that is each latent variable  $U^{(t)}$  is independent of  $U^{(t-2)}, \dots, U^{(1)}$ , given  $U^{(t-1)}$ . The resulting model is represented by the path diagram in Figure 1.

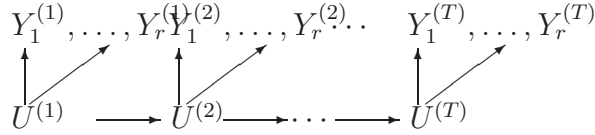


Figure 1: *Path diagram of the basic latent Markov model for multivariate data*

The model is characterized by three different types of parameters:

- the conditional response probabilities

$$\phi_{jy|u}^{(t)} = p(Y_j^{(t)} = y | U^{(t)} = u),$$

with  $j = 1, \dots, r$ ,  $t = 1, \dots, T$ ,  $u = 1, \dots, k$ , and  $y = 0, \dots, c_j - 1$ , which may be collected into the vector

$$\phi_{\mathbf{y}|u}^{(t)} = \prod_{j=1}^r \phi_{jy|u}^{(t)} = p(Y_1^{(t)} = y_1, \dots, Y_r^{(t)} = y_r | U^{(t)} = u);$$

- the initial probabilities

$$\pi_u = p(U^{(1)} = u),$$

with  $u = 1, \dots, k$ ;

- the transition probabilities

$$\pi_{u|v}^{(t|t-1)} = p(U^{(t)} = u | U^{(t-1)} = v),$$

with  $t = 2, \dots, T$ ,  $u, v = 1, \dots, k$ .

Note that all these probabilities do not depend on the specific sample unit. Moreover, it is possible to include a constraint on the transition probabilities corresponding to the hypothesis that the Markov chain is time homogeneous. Under this hypothesis, which is considered in the simulation study illustrated in Section 4, the transition probabilities do not depend on  $t$ , so as

$$\pi_{u|v}^{(t|t-1)} = \pi_{u|v}, \quad t = 2, \dots, T.$$

The number of free parameters of the multivariate LM model above is given by

$$\# \text{par} = k \underbrace{\sum_{j=1}^r (c_j - 1)}_{\phi_{jy|u}^{(t)}} + \underbrace{k - 1}_{\pi_u} + \underbrace{(T - 1)k(k - 1)}_{\pi_{u|v}^{(t|t-1)}}. \quad (1)$$

The probability mass function of the distribution of  $\mathbf{U}$  may be expressed as

$$p(\mathbf{U} = \mathbf{u}) = \pi_u \prod_{t=2}^T \pi_{u|v}^{(t|t-1)}$$

and the conditional distribution of  $\mathbf{Y}$  given  $\mathbf{U}$  is

$$p(\mathbf{Y} = \mathbf{y} | \mathbf{U} = \mathbf{u}) = \prod_{t=1}^T \phi_{\mathbf{y}|u}^{(t)} = \phi_{\mathbf{y}|u}^{(1)} \cdot \phi_{\mathbf{y}|u}^{(2)} \cdots \phi_{\mathbf{y}|u}^{(T)}.$$

Therefore, the manifest distribution  $p(\mathbf{Y} = \mathbf{y})$  of  $\mathbf{Y}$  follows

$$p(\mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{u}} \pi_u \pi_{u|v}^{(2|1)} \cdots \pi_{u|v}^{(T|T-1)} \phi_{jy|u}^{(1)} \cdots \phi_{jy|u}^{(T)}.$$



Note that computing  $p(\mathbf{Y} = \mathbf{y})$  involves all the possible  $k^T$  configurations of vector  $\mathbf{u}$ , that typically requires a considerable computational effort. In order to efficiently compute this probability we can use a forward recursion (Baum et al, 1970; Welch, 2003) for obtaining

$$q_{u,\mathbf{y}}^{(t)} = p(U^{(t)} = u, \mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(t)}).$$

In particular, the  $t$ -th iteration of this recursion, for  $t = 2, \dots, T$  consists of computing

$$q_{u,\mathbf{y}}^{(t)} = \sum_{v=1}^k q_{v,\mathbf{y}}^{(t-1)} \pi_{u|v}^{(t|t-1)} \phi_{\mathbf{y}|u}^{(t)} \quad u = 1, \dots, k$$

starting with  $q_{u,\mathbf{y}}^{(1)} = \pi_u \phi_{\mathbf{y}|u}^{(1)}$ , for  $t = 1$ . This recursion may be easily implemented by using matrix notation; see Bartolucci et al (2013) for details.

## 2.1 Likelihood inference

In an observed sample of  $n$  subjects, let  $n_{(\mathbf{y})}$  be the frequency of the observed response configuration  $\mathbf{y}$ , and assuming independence between the sample units, the model log-likelihood may be computed as

$$\ell(\boldsymbol{\theta}) = \sum_{\mathbf{y}} n_{(\mathbf{y})} \log[p(\mathbf{Y} = \mathbf{y})],$$

where  $\boldsymbol{\theta}$  is the vector of all model parameters arranged in a suitable way. The model log-likelihood may be maximized with respect to  $\boldsymbol{\theta}$  by using the Expectation-Maximization (EM) algorithm of Dempster et al (1977) which represents the main tool to estimate this class of models. This algorithm is based on the concept of *complete data*, which is represented by the pair  $(\mathbf{u}, \mathbf{y})$ , where  $\mathbf{u}$  denotes a realization of  $\mathbf{U}$ . Therefore, the

complete data log-likelihood is given by

$$\begin{aligned}\ell^*(\boldsymbol{\theta}) = & \sum_{j=1}^r \sum_{t=1}^T \sum_{u=1}^k \sum_{y=0}^{c-1} a_{juy}^{(t)} \log \phi_{jy|u}^{(t)} + \\ & + \sum_{u=1}^k b_u^{(1)} \log \pi_u + \sum_{t=2}^T \sum_{v=1}^k \sum_{u=1}^k b_{vu}^{(t)} \log \pi_{u|v}^{(t|t-1)}\end{aligned}$$

where  $a_{juy}^{(t)}$  corresponds to the frequency of subjects responding by  $y$  for the  $j$ -th response variable and belonging to latent state  $u$  at time  $t$ ,  $b_u^{(1)}$  is the frequency of subjects in latent state  $u$  at time 1, and  $b_{vu}^{(t)}$  corresponds to the frequency of subjects which move from latent state  $v$  to state  $u$  at time  $t$ .

Since the latent configuration for each subject is not known the EM maximizes the log-likelihood above by alternating the following two steps until convergence:

- **E-step:** compute the expected value of the above frequencies, given the observed data and the current value of the parameters, so as to obtain the expected value of  $\ell^*(\boldsymbol{\theta})$
- **M-step:** update  $\boldsymbol{\theta}$  by maximizing the expected value of  $\ell^*(\boldsymbol{\theta})$  obtained above; explicit solutions for  $\boldsymbol{\theta}$  estimation are available at this aim, see Bartolucci et al (2013).

The E-step of the algorithm involves the computation of the posterior probabilities  $f_{u|\mathbf{y}}^{(t)}$  and  $f_{u|v,\mathbf{y}}^{(t|t-1)}$ . Using the following backward recursion

$$\bar{q}_{v,\mathbf{y}}^{(t)} = p(\mathbf{Y}^{(t+1)}, \dots, \mathbf{Y}^{(T)} | U^{(t)} = v) = \sum_{u=1}^k \bar{q}_{u,\mathbf{y}}^{(t+1)} \pi_{u|v}^{(t+1|t)} \phi_{\mathbf{y}|u}^{(t+1)} \quad v = 1, \dots, k,$$

starting with  $\bar{q}_{v,\mathbf{y}}^{(T)} = 1$ , for  $t = 1, \dots, T$ , we have

$$f_{u|\mathbf{y}}^{(t)} = \frac{q_{v,\mathbf{y}}^{(t)} \bar{q}_{u,\mathbf{y}}^{(t)}}{P(\mathbf{Y} = \mathbf{y})}, \quad u = 1, \dots, k, \quad (2)$$

whereas, for  $t = 2, \dots, T$  and  $u, v = 1, \dots, k$ , we have

$$f_{u|v, \mathbf{y}}^{(t|t-1)} = \frac{q_{v, \mathbf{y}}^{(t-1)} \pi_{u|v}^{(t)} \phi_{\mathbf{y}|u}^{(t)} \bar{q}_{u, \mathbf{y}}^{(t)}}{P(\mathbf{Y} = \mathbf{y})}. \quad (3)$$

The recursions above may be implemented by using the matrix notation, as shown in Bartolucci (2006) and Bartolucci et al (2007).

### 3 The class of states selection criteria

As already discussed in Section 1, a crucial point in using LM models concerns the selection of the number of latent states  $k$ . When this number cannot be a priori defined, it is possible to rely on model selection criteria. In the following we illustrate the most common log-likelihood based information criteria together with classification based criteria which take the quality of classification into account.

#### 3.1 Log-likelihood based information criteria

The information criteria are based on indices that are, essentially, penalized versions of the maximum log-likelihood. The two most common criteria of this type are AIC and BIC. The first criterion, proposed by Akaike (1973), is a measure of the relative goodness of fit of a model, which describes the tradeoff between accuracy and complexity of the model. In particular, AIC is based on estimating the Kullback-Leibler distance between the true density and the estimated density, which focuses on the expected log-likelihood, and is defined on the basis of the following index

$$\text{AIC} = -2 \hat{\ell}(\boldsymbol{\theta}) + 2\#\text{par}. \quad (4)$$

For a given model,  $\hat{\ell}$  denotes the maximum of the log-likelihood of the LM model of interest and  $\#par$  denotes the number of free parameters as defined in (1). According to this criterion, the optimal number of latent states is that corresponding to the minimum value of the index in (4). In practice, we fit the LM model for increasing values of  $k$  until the index does not start to increase. Then, we select the previous  $k$  as the optimal number of latent states, which guarantees the best compromise between goodness-of-fit and model parsimony.

The BIC criterion of Schwarz (1978) is derived, for regular models, as an approximation to twice the log integrated likelihood (Kass and Raftery, 1995), using the Laplace method (Tierney and Kadane, 1986). From the asymptotic behavior of this approximation, the corresponding index may be defined as

$$\text{BIC} = -2 \hat{\ell}(\boldsymbol{\theta}) + \#par \log(n), \quad (5)$$

with  $\hat{\ell}$  and  $\#par$  defined as above. In certain settings, model selection based on BIC is roughly equivalent to model selection based on Bayes factors; see among others Kass and Raftery (1995). The number of latent states  $k$  to be selected is the one which corresponds to the minimum value of the index in (5). Usually, BIC leads to selecting a smaller number of latent states than the AIC criterion, since it is based on a more severe penalization. This difference may be relevant in complex model. In particular, the BIC criterion is expected to perform better as the amount of information increases with respect to the model complexity. In the LM literature, the same criteria have been used for model selection by Langeheine (1994), Langeheine and Van de Pol (1994), Magidson and Vermunt (2001), among many others. Finally, a comparison of their performance in connection with states selection of a univariate LM model may be found in Bartolucci

et al (2013), Ch.3. Moreover, comparisons between AIC and BIC criteria can be found in the literature of mixture models (McLachlan and Peel, 2000, Ch. 6), and in the HM literature for time series. From these studies, it emerges that BIC is usually preferable to AIC, as the latter tends to overestimate the number of states.

Among the variants of the AIC criterion existing in literature we also consider the criterion introduced by Bozdogan (1993). In particular, this criterion defines a more penalized version of the index in (4), on the basis of the results in Wolfe (1970), so as to obtain

$$\text{AIC}_3 = -2 \hat{\ell}(\boldsymbol{\theta}) + 3\#\text{par}, \quad (6)$$

in which the penalizing term 2 is substituted with 3.

On the other hand, the Consistent AIC criterion (CAIC), proposed by Bozdogan (1987), includes a penalizing term which also takes into account the sample size  $n$ , and is defined as

$$\text{CAIC} = -2 \hat{\ell}(\boldsymbol{\theta}) + \#\text{par}(\log(n) + 1). \quad (7)$$

Further to the above information criteria that are aimed at measuring the goodness of fit of a model, we also consider criteria that take into account the performance of the classification procedure, as outlined in the following.

### 3.2 Classification based information criteria

The criteria developed in the context of the classification likelihood approach, also known as complete data information criteria, are based on data augmentation, that is, the complete data as defined in Section 2.1. These criteria consider the following relation, that

was first showed by Hathaway (1986)

$$\ell^*(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \text{EN}, \quad (8)$$

see also Celeux and Soromenho (1996) and Biernacki and Govaert (1997), where EN is an entropy measure, which involves the posterior probabilities of component membership of each subject belonging to a specific group. Such entropy may be seen as a penalization term which is a measure of the ability of the model to provide a relevant partition of the data. More in detail, if the components are well separated, the posterior probabilities tend to define a partition of the data, assuming values close to 1. As a consequence, the entropy will be close to 0. The entropy measure cannot be directly used to assess the number of clusters since  $\ell(\boldsymbol{\theta})$  is an increasing function of  $k$  and has to be renormalized. With reference to the mixture models, Celeux and Soromenho (1996) proposed to consider the NEC criterion, which is expressed by

$$\text{NEC} = \frac{\text{EN}}{\hat{\ell}_k(\boldsymbol{\theta}) - \hat{\ell}_1(\boldsymbol{\theta})}, \quad k \geq 2, \quad (9)$$

where  $\hat{\ell}_k(\boldsymbol{\theta})$  is the maximum log-likelihood in case of a  $k$  components mixture and  $\hat{\ell}_1(\boldsymbol{\theta})$  is the maximum log-likelihood in case of a 1 component mixture. As also illustrated in Biernacki and Govaert (1997), NEC must assume small values to obtain a compromise between a good classification feature and a good description of the data. Then, the optimal number of components is the one that minimizes the index in (9). It is worth noting that the NEC criterion is not defined when  $k = 1$ ; to deal with this problem Biernacki et al (1999) proposed an empirical version of the NEC for Gaussian mixture. Usually it is convention to use  $\text{NEC} = 1$  for  $k = 1$ .

In extending NEC to LM models, the difficulty is in considering entropy based on the

posterior probabilities of all the possible configurations of latent states, given the observed data, for every sample unit. Therefore, this “true” entropy can be computed only when we have a reduced number of times occasions and latent states. In the context of HM models this measure is defined by Hernando et al (2005) as

$$\begin{aligned} \text{EN} &= - \sum_{u_1} \dots \sum_{u_T} f_{u_1, \dots, u_T | \mathbf{y}} \log(f_{u_1, \dots, u_T | \mathbf{y}}) = \\ &= - \sum_{u_1} \dots \sum_{u_T} f_{u_1 | \mathbf{y}}^{(1)} \cdot f_{u_2 | u_1, \mathbf{y}}^{(2|1)} \cdot \dots \cdot f_{u_t | u_{t-1}, \mathbf{y}}^{(t|t-1)} \cdot \dots \cdot f_{u_T | u_{T-1}, \mathbf{y}}^{(T|T-1)} \cdot \\ &\quad \cdot [\log(f_{u_1 | \mathbf{y}}^{(1)}) + \log(f_{u_2 | u_1, \mathbf{y}}^{(2|1)}) + \dots + \log(f_{u_t | u_{t-1}, \mathbf{y}}^{(t|t-1)}) + \dots + \log(f_{u_T | u_{T-1}, \mathbf{y}}^{(T|T-1)})] \end{aligned}$$

where  $f_{u | \mathbf{y}}^{(t)}$  and  $f_{u | v, \mathbf{y}}^{(t|t-1)}$  are defined in equations (2) and (3), respectively.

In the context of LM models, we may simplify the above equation for EN by formulating an approximated version that allows us to compute entropy also for any number of time occasions and latent states. More precisely, under the assumption that  $u^{(t)}$  are independent given  $\mathbf{Y}$ , we define EN as follows

$$\text{EN}_1 = - \sum_{u_1} \dots \sum_{u_T} f_{u | \mathbf{y}}^{(t)} \log(f_{u | \mathbf{y}}^{(t)}).$$

A possible renormalized variant of  $\text{EN}_1$  may also be expressed as

$$\text{EN}_2 = - \sum_{u_1} \dots \sum_{u_T} f_{u | \mathbf{y}}^{(t)} \log(f_{u | \mathbf{y}}^{(t)}) / T.$$

Therefore, we consider a NEC criterion relying on the “true” entropy based on the posterior probabilities of all possible configurations of latent states, and two different approximated versions, which may be computed for any number of time occasions and latent states. As an example we suppose to observe subjects at three occasions ( $T = 3$ ), then

the above criteria may be explicitly written as follows

$$\begin{aligned}
\text{EN} &= - \sum_u \sum_v \sum_z f_{u,v,z|\mathbf{y}} \log(f_{u,v,z|\mathbf{y}}) = \\
&= f_{z|v,\mathbf{y}}^{(3|2)} \cdot f_{v|u,\mathbf{y}}^{(2|1)} \cdot f_{u|\mathbf{y}}^{(1)} \cdot \\
&\quad \cdot [\log(f_{z|v,\mathbf{y}}^{(3|2)}) + \log(f_{v|u,\mathbf{y}}^{(2|1)}) + \log(f_{u|\mathbf{y}}^{(1)})] \\
\text{EN}_1 &= -[f_{u|\mathbf{y}}^{(1)} \cdot \log(f_{u|\mathbf{y}}^{(1)}) + f_{v|\mathbf{y}}^{(2)} \cdot \log(f_{v|\mathbf{y}}^{(2)}) + f_{z|\mathbf{y}}^{(3)} \cdot \log(f_{z|\mathbf{y}}^{(3)})] \\
\text{EN}_2 &= \frac{1}{3} \text{EN}_1
\end{aligned}$$

According with the entropy measures defined above we consider three different versions of the NEC criterion where the first one is based on the “true” entropy, as defined in (9), and the other two versions are expressed as

$$\begin{aligned}
\text{NEC}_1 &= \frac{\text{EN}_1}{\hat{\ell}(\boldsymbol{\theta}) - \hat{\ell}_1(\boldsymbol{\theta})}, \quad k \geq 2; \\
\text{NEC}_2 &= \frac{\text{EN}_2}{\hat{\ell}(\boldsymbol{\theta}) - \hat{\ell}_1(\boldsymbol{\theta})}, \quad k \geq 2.
\end{aligned}$$

Among other criteria which take the quality of classification into account, we also consider the CLC criterion, proposed by Biernacki and Govaert (1997) in the mixture context, which uses the relation in (8) to define the following index

$$\text{CLC} = -2\hat{\ell}(\boldsymbol{\theta}) + 2 \text{EN}.$$

Moreover, Biernacki et al (1998) suggested an alternative information criterion based on the complete data likelihood named as Integrated Classification Likelihood criterion (ICL). The same authors also proposed an approximated version of the ICL using BIC (Biernacki et al, 2000). In particular, McLachlan and Peel (2000) referred to this approx-



imated version as ICL-BIC and showed that may be computed as

$$\text{ICL-BIC} = \text{BIC} + 2 \text{ EN},$$

in which the term  $2 \text{ EN}$  represents a kind of penalization for poorly separated clusters; see also Li (2005).

As already discussed in Section 1, although the above information criteria are widely used and their performance are studied in the context of finite mixture, LC, and HM models, there is still a lack in the literature about their comparison in the context of LM models. In the following section we set an experimental design to compare the performance of these criteria in the context of multivariate LM model, in order to choose the optimal number of latent states.

## 4 Simulation study

We illustrate the results obtained by a Monte Carlo simulation study aimed at comparing the performance of the following indices for states selection:

- Log-likelihood based criteria: AIC, CAIC, AIC<sub>3</sub>, BIC;
- Classification based criteria: CLC, ICL-BIC, NEC, NEC<sub>1</sub>, NEC<sub>2</sub>.

More in detail, we simulate 100 samples with a given size  $n$  ( $n = 250, 500$ ) and coming from an LM model, characterized by  $r$  ( $r = 1, 3, 5$ ) binary ( $y = 0, 1$ ) response variables observed in  $T = 5$  time occasions,  $k$  ( $k = 2, 3$ ) latent states, and given values of initial probabilities  $\pi_u$ , transition probabilities  $\pi_{u|v}^{(t|t-1)}$ , and conditional response probabilities  $\phi_{jy|u}^{(t)}$ . All analyses are implemented in R software (the code is available upon request by

authors). In all cases, the strategy adopted to choose the number of latent states is the same: we fit the LM model with increasing  $k$  values and, then, the value just before the first increasing criterion index is taken as optimal number of latent states.

We first consider a scenery (scenery 1) based on  $n = 250$  individuals belonging to  $k = 2$  latent states and observed on  $T = 5$  time occasions. Moreover, we suppose equal initial probabilities, that is  $\pi_1 = 0.5 = \pi_2$ , and denoting by  $\mathbf{\Pi}$  the transition probabilities matrix with elements  $\pi_{u|v}$ , under the time homogeneous assumption, we consider

$$\mathbf{\Pi} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

We also assume alternatively  $r = 1, 3, 5$  binary response variables with the following matrix  $\mathbf{\Phi}$  of the conditional response probabilities  $\phi_{jy|u}^{(t)}$

$$\mathbf{\Phi} = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}.$$

Table 1 shows the relative frequencies of the number  $k$  of components chosen by each of the considered criteria, both in univariate case ( $r = 1$ ) and in multivariate cases ( $r = 3$  and  $r = 5$ ).

In the univariate case, all log-likelihood based criteria perform very well, whereas the performance of classification based criteria is very bad: they tend to underestimate  $k$  in almost all cases. Instead, in the multivariate cases ( $r = 3$  and  $r = 5$ ) the classification based criteria improve considerably their performance, being the selected number of latent states equal to 2 in almost all cases. We only observe a worsening of AIC, which tends to overestimate the right number of latent states in 17 and 29 cases out of 100, for  $r = 3$  and for  $r = 5$  respectively.

Table 1: Relative frequencies of  $k$  chosen on the basis of several criteria (scenery 1,  $n = 250$ )

$k$	BIC	AIC	AIC <sub>3</sub>	CAIC	NEC	NEC <sub>1</sub>	NEC <sub>2</sub>	CLC	ICL-BIC
$r = 1$									
1	0.00	0.00	0.00	0.00	<b>1.00</b>	<b>0.95</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
2	<b>1.00</b>	<b>0.99</b>	<b>1.00</b>	<b>1.00</b>	0.00	0.05	0.00	0.00	0.00
3	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	<b>1.00</b>	<b>0.83</b>	<b>1.00</b>	<b>1.00</b>	<b>0.99</b>	<b>1.00</b>	<b>1.00</b>	<b>0.98</b>	<b>1.00</b>
3	0.00	0.14	0.00	0.00	0.00	0.00	0.01	0.00	0.00
4	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	<b>1.00</b>	<b>0.71</b>	<b>0.97</b>	<b>1.00</b>	<b>0.99</b>	<b>1.00</b>	<b>1.00</b>	<b>0.93</b>	<b>1.00</b>
3	0.00	0.26	0.03	0.00	0.01	0.00	0.00	0.05	0.00
4	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00

An alternative scenery (scenery 2) is then considered, which differs from scenery 1 for lower values of state persistence probabilities given by

$$\mathbf{\Pi} = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}.$$

All the other elements are the same than those considered in scenery 1. Results are shown in Table 2.

With respect to scenery 1 we note several differences. In the univariate case, the behavior of both BIC and CAIC gets worse: BIC leads to select the true value  $k = 2$  in less than 50% of cases and CAIC in 37% of cases, whereas in the remaining simulations

Table 2: Relative frequencies of  $k$  chosen on the basis of several criteria (scenery 2,  $n = 250$ )

$k$	BIC	AIC	AIC <sub>3</sub>	CAIC	NEC	NEC <sub>1</sub>	NEC <sub>2</sub>	CLC	ICL-BIC
$r = 1$									
1	<b>0.52</b>	0.00	0.10	<b>0.63</b>	<b>1.00</b>	<b>1.00</b>	<b>0.99</b>	<b>1.00</b>	<b>1.00</b>
2	0.48	<b>0.98</b>	<b>0.90</b>	0.37	0.00	0.00	0.01	0.00	0.00
3	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	<b>0.88</b>	<b>0.92</b>	0.00	<b>0.88</b>	<b>0.95</b>
2	<b>1.00</b>	<b>0.83</b>	<b>0.98</b>	<b>1.00</b>	0.10	0.07	<b>0.96</b>	0.10	0.04
3	0.00	0.16	0.02	0.00	0.01	0.01	0.04	0.01	0.01
4	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	<b>1.00</b>	<b>0.77</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
3	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00

both of them underestimate  $k$ . A slightly worsening is observed also for AIC<sub>3</sub>, which in the 10% of simulated models chooses only one latent state. With  $r = 3$  responses, on one hand BIC, AIC<sub>3</sub>, CAIC and NEC<sub>2</sub> considerably improve their performance and, on the other hand, AIC tends to overestimate  $k$  in 17 cases out of 100, obtaining in both situations values similar to those of scenery 1. However, the classification based criteria other than NEC<sub>2</sub> improve just a little and continue to underestimate  $k$  in the main part of cases, showing a really different behavior with respect to scenery 1. Similarly to scenery 1, with  $r = 5$  all the considered criteria present an optimal behavior (with the exception of AIC, which overestimates  $k$  in 23 cases out of 100).

Another scenery (scenery 3) is then considered, which differs from scenery 1 for a greater uncertainty in the allocation of the observations to the latent states, being the conditional response probabilities matrix given by

$$\Phi = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}.$$

All the other elements are the same than scenery 1. Results are shown in Table 3.

Table 3: Relative frequencies of  $k$  chosen on the basis of several criteria (scenery 3,  $n = 250$ )

$k$	BIC	AIC	AIC <sub>3</sub>	CAIC	NEC	NEC <sub>1</sub>	NEC <sub>2</sub>	CLC	ICL-BIC
$r = 1$									
1	0.35	0.01	0.02	<b>0.53</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
2	<b>0.65</b>	<b>0.98</b>	<b>0.97</b>	0.47	0.00	0.00	0.00	0.00	0.00
3	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	<b>0.99</b>	<b>1.00</b>	0.08	<b>0.99</b>	<b>1.00</b>
2	<b>1.00</b>	<b>0.79</b>	<b>1.00</b>	<b>1.00</b>	0.01	0.00	<b>0.88</b>	0.01	0.00
3	0.00	0.18	0.00	0.00	0.00	0.00	0.03	0.00	0.00
4	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
$r = 5$									
1	0.00	0.00	0.000	0.00	0.285	<b>0.770</b>	0.000	0.285	<b>0.550</b>
2	<b>1.00</b>	<b>0.78</b>	<b>0.995</b>	<b>1.00</b>	<b>0.590</b>	0.220	<b>0.980</b>	<b>0.585</b>	0.445
3	0.00	0.205	0.005	0.00	0.030	0.005	0.015	0.035	0.005
4	0.00	0.01	0.00	0.00	0.070	0.005	0.005	0.070	0.000
5	0.00	0.005	0.00	0.00	0.025	0.000	0.000	0.025	0.000

With respect to scenery 1, in presence of  $r = 1$  response variables the behavior of BIC and CAIC is not very satisfactory, because they tend to underestimate  $k$  in 35% and 53% of cases. Concerning the classification based criteria, we note a significant deterioration

of their behavior both in univariate case and in multivariate cases. Only  $\text{NEC}_2$  presents a satisfactory performance, being the correct number of  $k$  selected in 88 cases out of 100 with  $r = 3$  and in 98 cases out of 100 with  $r = 5$  (and overestimated in the remaining cases). Instead, the remaining criteria lead to choose  $k = 1$  in almost all cases when  $r = 3$ , improving just a little when  $r = 5$ . More precisely, in this last case, NEC and CLC allow us to select the right number of  $k$  in the 59% of cases, whereas they underestimate  $k$  in 28.5% of cases. Moreover, ICL-BIC leads to select  $k = 1$  for the 55% of simulated models and  $\text{NEC}_1$  for the 77% of them.

The three above described sceneries are then replicated by increasing the number of observations from  $n = 250$  to  $n = 500$ , all the other things being constant. Results are shown in Tables 4, 5, and 6 for sceneries 1, 2, and 3 respectively.

By increasing the number of observations we note a considerable improvement of performances of BIC and CAIC in the univariate cases of sceneries 2 and 3, whereas the behavior of the other criteria, especially of classification based criteria, is unchanged. Rather, in case of scenery 3, when  $r = 5$  response variables are considered, the behavior of  $\text{NEC}_1$  and ICL-BIC gets worse, being  $k = 2$  selected in 10% and 43% of cases for  $n = 500$  against 22% and 44.5% of cases for  $n = 250$ .

To conclude, we also consider two further sceneries (sceneries 4 and 5) characterized by  $n = 500$  individuals and  $k = 3$  latent states. We also suppose  $T = 5$ , equal initial probabilities, that is  $\pi_1 = \pi_2 = \pi_3 = 1/3$ , conditional response probabilities matrix equal to

$$\Phi = \begin{pmatrix} 0.9 & 0.1 & 0.7 \\ 0.1 & 0.9 & 0.3 \end{pmatrix},$$

Table 4: Relative frequencies of  $k$  chosen on the basis of several criteria (scenery 1,  $n = 500$ )

$k$	BIC	AIC	AIC <sub>3</sub>	CAIC	NEC	NEC <sub>1</sub>	NEC <sub>2</sub>	CLC	ICL-BIC
$r = 1$									
1	0.00	0.00	0.00	0.00	<b>1.00</b>	<b>1.00</b>	<b>0.98</b>	<b>1.00</b>	<b>1.00</b>
2	<b>1.00</b>	<b>0.99</b>	<b>1.00</b>	<b>1.00</b>	0.00	0.00	0.02	0.00	0.00
3	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	<b>1.00</b>	<b>0.87</b>	<b>1.00</b>	<b>1.00</b>	<b>0.99</b>	<b>1.00</b>	<b>1.00</b>	<b>0.99</b>	<b>1.00</b>
3	0.00	0.13	0.00	0.00	0.01	0.00	0.00	0.01	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	<b>1.00</b>	<b>0.78</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
3	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

and the following transition probabilities (under the time homogeneous assumption)

$$\mathbf{\Pi} = \begin{pmatrix} 0.90 & 0.05 & 0.05 \\ 0.05 & 0.90 & 0.05 \\ 0.05 & 0.05 & 0.90 \end{pmatrix},$$

in the first case (scenery 4) and

$$\mathbf{\Pi} = \begin{pmatrix} 0.70 & 0.15 & 0.15 \\ 0.15 & 0.70 & 0.15 \\ 0.15 & 0.15 & 0.70 \end{pmatrix},$$

in the second case (scenery 5).

With respect to the cases with  $k = 2$  latent states, we now observe a very poor performance of all criteria in the univariate case ( $r = 1$ ): log-likelihood based criteria

Table 5: Relative frequencies of  $k$  chosen on the basis of several criteria (scenery 2,  $n = 500$ )

$k$	BIC	AIC	AIC <sub>3</sub>	CAIC	NEC	NEC <sub>1</sub>	NEC <sub>2</sub>	CLC	ICL-BIC
$r = 1$									
1	0.05	0.00	0.01	0.08	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
2	<b>0.95</b>	<b>0.99</b>	<b>0.99</b>	<b>0.92</b>	0.00	0.00	0.00	0.00	0.00
3	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	<b>0.97</b>	<b>0.98</b>	0.00	<b>0.97</b>	<b>0.98</b>
2	<b>1.00</b>	<b>0.74</b>	<b>0.99</b>	<b>1.00</b>	0.03	0.02	<b>1.00</b>	0.03	0.02
3	0.00	0.23	0.01	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	<b>1.00</b>	<b>0.74</b>	<b>0.98</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
3	0.00	0.17	0.02	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

lead to select  $k = 2$  and classification based criteria lead to select  $k = 1$ , in almost all cases (Tables 7 and 8). Instead, in presence of  $r = 3$  response variables, the behavior of log-likelihood based criteria gets better, especially under scenery 4. Indeed, Table 7 shows that with AIC and AIC<sub>3</sub> the right number of latent states is chosen in 91 and 98 cases out of 100 respectively, whereas with BIC and CAIC the percentages of right choices are reduced to 59% and 39% respectively, being  $k$  underestimated in the remaining cases. On the other hand, under scenery 5 (Table 8), which refers to a situation with lower latent states persistence probabilities, the improvement is far from clear:  $k = 3$  is selected by AIC in 75% of cases and by AIC<sub>3</sub> in 35%, whereas BIC and CAIC lead to choose regularly



Table 6: Relative frequencies of  $k$  chosen on the basis of several criteria (scenery 3,  $n = 500$ )

$k$	BIC	AIC	AIC <sub>3</sub>	CAIC	NEC	NEC <sub>1</sub>	NEC <sub>2</sub>	CLC	ICL-BIC
$r = 1$									
1	0.01	0.00	0.00	0.03	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
2	<b>0.99</b>	<b>1.00</b>	<b>1.00</b>	<b>0.97</b>	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	<b>1.00</b>	<b>1.00</b>	0.04	<b>1.00</b>	<b>1.00</b>
2	<b>1.00</b>	<b>0.86</b>	<b>0.97</b>	<b>1.00</b>	0.00	0.00	<b>0.96</b>	0.00	0.00
3	0.00	0.12	0.03	0.00	0.00	0.00	0.000	0.00	0.00
4	0.00	0.01	0.00	0.00	0.00	0.00	0.000	0.00	0.00
5	0.00	0.01	0.00	0.00	0.00	0.00	0.000	0.00	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.32	<b>0.90</b>	0.00	0.32	<b>0.57</b>
2	<b>1.00</b>	<b>0.67</b>	<b>1.00</b>	<b>1.00</b>	<b>0.63</b>	0.10	<b>1.00</b>	<b>0.63</b>	0.43
3	0.00	0.27	0.00	0.00	0.01	0.00	0.00	0.01	0.00
4	0.00	0.04	0.00	0.00	0.03	0.00	0.00	0.03	0.00
5	0.00	0.02	0.00	0.00	0.01	0.00	0.00	0.01	0.00

$k = 2$ . Finally, with  $r = 5$  responses we observe satisfactory performances of log-likelihood based criteria (although AIC overestimates  $k$  in 15% of cases) in case of scenery 4 (Table 7), but, again under scenery 5, results are not very satisfactory, because BIC and CAIC performs well in only 76% and 52% of cases. The behavior of classification based criteria is definitely disappointing under both sceneries and both with  $r = 3$  and  $r = 5$ .

## 5 Conclusions

In this paper we investigated about a typical issue characterizing some latent variable models, such as latent class models or hidden Markov (HM) models, consisting in the

Table 7: Relative frequencies of  $k$  chosen on the basis of several criteria (scenery 4)

$k$	BIC	AIC	AIC <sub>3</sub>	CAIC	NEC	NEC <sub>1</sub>	NEC <sub>2</sub>	CLC	ICL-BIC
$r = 1$									
1	0.00	0.00	0.00	0.00	<b>1.00</b>	<b>1.00</b>	0.00	<b>1.00</b>	<b>1.00</b>
2	<b>1.00</b>	<b>0.93</b>	<b>0.99</b>	<b>1.00</b>	0.00	0.00	<b>1.00</b>	0.00	0.00
3	0.00	0.07	0.01	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.41	0.01	0.02	<b>0.62</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
3	<b>0.59</b>	<b>0.91</b>	<b>0.98</b>	0.39	0.00	0.00	0.00	0.00	0.00
4	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
3	<b>1.00</b>	<b>0.85</b>	<b>0.99</b>	<b>1.00</b>	0.00	0.00	0.00	0.00	0.00
4	0.00	0.13	0.01	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00

choice about the number of mixture components (i.e, latent classes or latent states). More precisely, we focused on the selection of latent states in univariate and multivariate latent Markov (LM) models for longitudinal data. We firstly illustrated the assumptions and the structure of LM model, giving some hints about the maximization of log-likelihood on the basis of an EM algorithm. Then, we described some of the most well-known model selection criteria used in the context of mixture models, distinguishing between log-likelihood based criteria (i.e., AIC, AIC<sub>3</sub>, CAIC, BIC) and classification based criteria (NEC, CLC, ICL-BIC). Concerning this latter type of criteria, we gave some emphasis to the problem of properly defining the entropy in case of LM models. Relying on the case of

Table 8: Relative frequencies of  $k$  chosen on the basis of several criteria (scenery 5)

$k$	BIC	AIC	AIC <sub>3</sub>	CAIC	NEC	NEC <sub>1</sub>	NEC <sub>2</sub>	CLC	ICL-BIC
$r = 1$									
1	0.00	0.00	0.00	0.00	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
2	<b>1.00</b>	<b>0.97</b>	<b>1.00</b>	<b>1.00</b>	0.00	0.00	0.00	0.00	0.00
3	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	<b>0.99</b>	0.15	<b>0.65</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
3	0.01	<b>0.75</b>	0.35	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.24	0.00	0.01	0.48	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>	<b>1.00</b>
3	<b>0.76</b>	<b>0.86</b>	<b>0.99</b>	<b>0.52</b>	0.00	0.00	0.00	0.00	0.00
4	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

HM models, we observed the possibility of computing the exact entropy only in presence of a reduced number of time occasions and latent states. Therefore, we also proposed two variants (named NEC<sub>1</sub> and NEC<sub>2</sub>) that can be easily computed also for any number of time occasions and latent states.

On the basis of some Monte Carlo simulations, we compared the performance of log-likelihood and classification based criteria for the latent states selection in univariate and multivariate LM models. Generally speaking, lower values of persistence probabilities in a same state and/or a greater uncertainty in the allocation of the observations to the latent states complicate the task of latent states selection procedures, leading to a generally worse performance of the adopted criteria. Instead, the number of observations does not

play a relevant role.

Concerning the specific criteria, we observed that those based on the log-likelihood present a better general behavior with respect to those based on classification, even if AIC tends to overestimate the correct number of latent states, especially in multivariate cases. The classification based criteria tend to underestimate the true number of latent states, mainly for the univariate case, whereas their performance improves by increasing the number of observed response variables. We also observed a significant better behavior for  $NEC_2$  with respect to the other classification based criteria. Finally, by increasing the number of latent states the performance of all considered criteria gets worse, mainly in the univariate case. We conclude outlining that the results we obtained are coherent with those observed in the literature about HM models (see, among others, Costa and De Angelis, 2010).

Concerning further developments of the present work, we intend to extend the simulation study to LM models with covariates. We also intend to rely on the most recent advances in the context of HM models for a different formulation of the entropy that takes into account the tendency of traditional formulation to overestimate the uncertainty in these type of models (Durand and Guédon, 2012).

## References

Akaike H (1973) Information theory and an extension of the Maximum Likelihood principle. In: Petrov B, Csaki F (eds) Second International Symposium on Information Theory, Akademiai Kiado, Budapest, pp 267–281

- Bartolucci F (2006) Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *Journal of the Royal Statistical Society, series B* 68:155–178
- Bartolucci F, Colombi R, Forcina A (2007) An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statistica Sinica* 17:692–711
- Bartolucci F, Farcomeni A, Pennoni F (2013) *Latent Markov Models for Longitudinal Data*. Chapman & Hall/CRC, Boca Raton, FL
- Baum LE, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics* 41:164–171
- Biernacki C, Govaert G (1997) Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics* 29:451–457
- Biernacki C, Govaert G (1999) Choosing models in model-based clustering and discriminant analysis. *J Statistical Computation and Simulation* 14:49–71
- Biernacki C, Celeux G, Govaert G (1998) Assessing a mixture model for clustering with the integrated classification likelihood. PhD thesis, Institut National de Recherche en informatique et en automatique
- Biernacki C, Celeux G, Govaert G (1999) An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters* 20:267–272

- Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22:719–725
- Bozdogan H (1987) Model selection and Akaike’s Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 52:345–370
- Bozdogan H (1993) Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the Inverse-Fisher information matrix. In: Opitz O, Lausen B, Klar R (eds) *Information and Classification, Concepts, Methods and Applications*, Springer, Berlin, pp 40–54
- Celeux G, Durand JB (2008) Selecting hidden Markov model state number with cross-validated likelihood. *Comput Stat* 23:541–564
- Celeux G, Soromenho G (1996) An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification* 13:195–212
- Costa M, De Angelis L (2010) Model selection in hidden Markov models: a simulation study. *Quaderni di Dipartimento 7*, Department of Statistics, University of Bologna
- Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39:1–38
- Dias J (2006) Model selection for the binary latent class model: A Monte Carlo simulation. In: Batagelj V, Bock HH, Ferligoj A, Iberná A (eds) *Data Science and Classification*, Springer Berlin Heidelberg, pp 91–99

- Durand JB, Guédon Y (2012) Localizing the latent structure canonical uncertainty: entropy profiles for hidden Markov models. Tech. rep., Research Report 7896, Project-Teams Mistis and Virtual Plants
- Fraley C, Raftery A (2002) Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97:611–631
- Hathaway RJ (1986) Another interpretation of the EM algorithm for mixture distributions. *Statistics & Probability Letters* 4:53 – 56
- Hernando D, Crespi V, Cybenko G (2005) Efficient computation of the hidden Markov model entropy for a given observation sequence. *IEEE Transactions on Information Theory* 51:2681–2685
- Kass RE, Raftery AE (1995) Bayes factors and model uncertainty. *Journal of the American Statistical Association* 90:773–795
- Langeheine R (1994) Latent variables Markov models. In: von Eye A, Clogg C (eds) *Latent variables analysis: Applications for developmental research*, Sage, Thousand Oaks, CA, pp 373–395
- Langeheine R, Van de Pol F (1994) Discrete time mixed Markov latent class models. In: Dale A, RB D (eds) *Analyzing social and political change. A casebook of methods.*, London: Sage., pp 171–197
- Li J (2005) Clustering based on a multilayer mixture model. *Journal of Computational and Graphical Statistics* 14:547–568

- Lin TH, Dayton CM (1997) Model selection information criteria for non-nested latent class models. *Journal of Educational and Behavioral Statistics* 22:249–264
- Magidson J, Vermunt JK (2001) Latent class factor and cluster models, bi-plots and related graphical displays. *Sociological Methodology* 31:223–264
- McLachlan G, Peel D (2000) *Finite Mixture Models*. Wiley
- Nylund K, Asparouhov T, Muthén B (2006) Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling* 14:535–569
- Schwarz G (1978) Estimating the dimension of a model. *The Annals of Statistics* 6:461–464
- Tierney L, Kadane JB (1986) Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association* 81:82–86
- Welch LR (2003) Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter* 53:1–13
- Wiggins L (1973) *Panel analysis. Latent probability models for attitude and behavior processes*. Elsevier, New York, US
- Wolfe JH (1970) Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research* 5:329–350
- Zucchini W, MacDonald IL (2009) *Hidden Markov Models for time series: an introduction using R*. Springer-Verlag, New York